Two-level non-parametric scaling for dichotomous data

Tom A.B. Snijders
ICS/Department of Statistics and Measurement Theory
University of Groningen [*]

Abstract

It is relevant to extend the existing single-level scaling methods to two-level designs. Examples are the scaling of teachers on the basis of their pupils' responses, or scaling neighborhoods on the basis of responses by inhabitants. A non-parametric approach is convenient because it requires few assumptions and leads to easy calculations.

This paper considers a two-level situation where the objects to be scaled are the higher level units; nested within each object are lower level units, called 'subjects'; a set of dichotomous items is administered to each subject. A two-level version is elaborated of the non-parametric scaling method first proposed by Mokken (1971). The probabilities of positive responses to the items are supposed to be increasing functions of the value on a latent trait; this value is composed of a subject-dependent value and a deviation from this value due to the object and the subject-object interaction. This situation may be viewed as one with strictly parallel tests that are defined by the objects.

Loevinger H coefficients are defined to assess the consistency of responses within, but also between objects. The availability of parallel tests is used to calculate coefficient alpha to assess the reliability of the scale.

Keywords: Multi-level models, item response theory, non-parametric scaling, reliability, parallel tests, ecometrics.

[*] Department of Statistics and Measurement Theory, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, email t.a.b.snijders@ppsw.rug.nl, http://stat.gamma.rug.nl/snijders/ .

1 Introduction

In item response theory we are accustomed to situations where the subjects who provide the responses are also the objects which are to be measured; more generally, where for each individual to be scaled exactly one set of item responses is available. There are situations, however, where these roles are fulfilled by distinct entitities, as is demonstrated by the following examples.

1. Pupils are asked questions about their teachers, with the purpose to assess the teachers (as perceived by the pupils) on a certain dimension.

2. Employees are asked to respond to a questionnaire about the department they work in, in order to scale the departments with respect to working climate.

3. Social settings such as neigborhoods are assessed by a survey of multiple informants (e.g., inhabitants). The answers given by the informants are aggregated to provide an assessment of the social setting (Raudenbush and Sampson, 1998, use the term *ecometrics* for such an approach).

4. In a study of personal networks, for each respondent ('ego') in a survey, a list is made of the other persons ('alters') to whom the respondent is related according to certain criteria. A number of characteristics of the alters are collected in order to characterize ego's social environment (Jansson and Spreen, 1998).

The *objects* to be measured in these examples are, respectively, teachers, departments, social settings, and respondents. For each object a set of multiple measurements is available: multiple pupils, employees, informants, and alters. These units will be called *subjects*, although this word may be unexpected in the fourth example. Each of the multiple measurements is supposed to yield a vector of responses to the same set of *items*. This defines a three-level nesting structure: responses denoted Y_{ijk} to items k are nested within subjects j, which in their turn are nested within objects i. Equivalently, this can be regarded as a multivariate two-level nesting structure, where subjects are nested within objects and each subject provides a vector of responses $Y_{ij} = (Y_{ij1}, ..., Y_{ijK})$.

The purpose of the analysis is to give scale values to the objects, while the subjects are regarded as parallel tests of these objects. It will be assumed that the item responses Y_{ijk} are dichotomous, scored as 1 ('positive', 'correct') and 0 ('negative', 'false'). The statistical approach elaborated here is a non-parametric latent trait model in the sense that a one-dimensional latent trait is assumed to govern (stochastically) the item responses, and the assumptions about how the response probabilities depend on the latent trait are of a qualitative nature only and do not restrict these probabilities to a particular mathematical function. Relevant questions are the following:

- What is, in this two-level design, a suitable non-parametric definition of the relation between latent trait and observed responses?
- How to assess empirically whether indeed there is a one-dimensional latent variable which underlies the responses to the items?
- How can the objects be scored on this latent dimension?
- What is the reliability of the resulting scale?

These questions will be answered by elaborating a two-level version of the non-parametric scaling method first proposed by Mokken (1971) and explained and elaborated by Mokken and Lewis (1982), Mokken (1997), and Molenaar and Sijtsma (200?). Alternative parametric models are also available for this two-level scaling design, as will be mentioned in the Discussion.

The Mokken scaling model is an attractive non-parametric model for unidimensional cumulative scaling. This model can be summarized as follows. Subjects, to be indexed by the letter i, respond to dichotomously scored items indexed by the letter k. The response (scored 0 or 1) of object i to item k is denoted Y_{ik}. It is assumed that a latent trait θ exists for which each person has a value, denoted θ_i, determining the probability distribution of i's vector of responses. Further it is assumed that for each item k, there exists a non-decreasing function $p_k(.)$ such that the probability of a positive response, given the latent trait value θ_i, is

$$P\,\{Y_{ik} = 1 \,|\, \theta_i\} \;\; = \;\; p_k(\theta_i) \; . \tag{1}$$

These functions $p_k(.)$ are called tracelines, or item characteristic curves. Furthermore, the Mokken model assumes local stochastic independence: conditional on the value of the latent trait, the responses of the subject are outcomes of independent random variables. Important questions in the application of this scaling method are the scalability of a given set of items, the selection of a scalable subset from a given set of items, and the attribution of scores to subjects. Procedures for data analysis according to this model are described in the literature mentioned above, and implemented in the program MSP.

2 A two-level model for non-parametric scaling of dichotomous data

As stated above, Y_{ijk} denotes the response of subject j to item k with regard to object i. The set of items scored for each object-subject combination is assumed to be constant, while the number of subjects may vary between objects. The number of subjects providing responses with regard to object i is denoted n_i.

The model elaborated for this situation is a cumulative scaling model governed by a uni-dimensional latent trait θ. We assume that for each object i there is a value θ_i on the latent trait, and to each subject j nested within object i there corresponds a deviation δ_{ij} from this value. Combined, this yields the combined value $\theta_i + \delta_{ij}$ as the resultant value on the latent trait for the combination of subject j and object i. The deviation δ_{ij} may be considered as a (random) subject effect together with object by subject interaction. Conditional on these latent trait values, we assume stochastic independence of the responses to the different items and subjects. We also assume the existence of tracelines $p_k(.)$ that are specific to the items but do not depend on object or subject, and that are non-decreasing, and give the probabilities of positive responses:

$$P\{Y_{ijk} = 1 \,|\, \theta_i,\, \delta_{ij}\} \;=\; p_k(\theta_i + \delta_{ij}) \; . \tag{2}$$

Further, we assume that all subjects providing data for the objects are a random sample from some population of subjects. The parallel tests therefore are independent, conditionally on θ_i. For the statistical model this means that the δ_{ij} are independent and identically distributed random variables.

In terms of multilevel analysis (as treated by Bryk and Raudenbush, 1992; Goldstein, 1995; Snijders and Bosker, 1999), this may be regarded as a model where the random parts at levels two and three are composed of random intercepts without random slopes.

In Section 5 we treat the question of defining object scores in order to *scale* the objects. Scaling object i amounts in this model to estimating a suitable monotone increasing function of θ_i. (This function is defined below by (21); in this non-parametric framework the latent parameter θ_i itself is not identifiable, because no assumptions are made as to the shape of the tracelines; also see Mokken, 1971).

When $\text{var}(\theta_i)$ is large compared to $\text{var}(\delta_{ij})$, then the differences between the subjects hardly matter compared to the differences between the objects. This is desirable from the point of view of scaling the objects. When, on the contrary, $\text{var}(\theta_i)$ is small compared to $\text{var}(\delta_{ij})$, then the probabilities of the two responses to the items are determined by the subjects rather than by the objects, leading to a comparatively unreliable test (unless the number of subjects per object is large).

It follows from this model that the marginal probability of a positive response for a randomly drawn subject to object i is

$$\pi_k(\theta_i) \;=\; P\{Y_{ijk} = 1 \,|\, \theta_i\} \;=\; E_\delta \, p_k(\theta_i + \delta_{ij}) \;, \tag{3}$$

where the expectation E_δ refers to the distribution of δ_{ij}. The function $\pi_k(.)$ inherits from the function $p_k(.)$ the property of being monotone non-decreasing. (The property of double monotonicity, defined in the following section, is also inherited by $\pi_k(.)$ from $p_k(.)$.) However, the function $\pi_k(.)$ will be flatter than the original traceline $p_k(.)$, because of the process of averaging with respect to the distribution of δ_{ij}.

The questions concerning this model treated in this paper are the following:

- Model adequacy: do the items together form a cumulative scale which is suitable for scaling the objects? This is investigated on the basis of Loevinger scalability coefficients.
- The definition of scores for objects.
- The reliability of the resulting scale.

3 Scalability coefficients

In the usual form of Mokken scaling, scalability coefficients are defined as follows. These coefficients are based on a further assumption, namely, the assumption of *double monotonicity*: the items k are ordered in such a way, that

$$p_1(\theta) \le p_2(\theta) \le \ldots \le p_m(\theta) \qquad \text{for all } \theta,$$

where K is the number of items. For item pairs k, k' with $k < k'$, the responses Y_{ik}, Y_{ik}, are said to form an *error* if $Y_{ik} = 1$ while $Y_{ik'} = 0$, contradicting the expected ordering. For the definition of the scalability coefficients it is also assumed that objects have been drawn at random from some population, or that (more loosely) the θ_i values for the objects can be considered to be representative for some population. This assumption is needed because the scalability coefficients are defined relative to the *marginal* joint probability distribution of the item scores, i.e., the distribution of $(Y_{i1}, Y_{i2}, \ldots, Y_{iK})$ for a randomly drawn object; these coefficients depend jointly on the items and the population of objects. Loevinger's H coefficients of scalability can now be defined (see Mokken, 1971 and Mokken and Lewis, 1982) with respect to each item pair ($k < k'$) as one minus the ratio of the probability of an error response for the item pair to the same probability under the null model of independent item responses,

$$H_{kk'} \;=\; 1 - \frac{P\{Y_{ik} = 1,\, Y_{ik'} = 0\}}{P\{Y_{ik} = 1\}\, P\{Y_{ik'} = 0\}} \;.$$

In the case of a perfect Guttman scale this scalability coefficient equals 1, in the case of independent (unrelated) items this coefficient equals 0. A similar coefficient can be defined for each item, by considering the total number of errors; note that for item k, pattern $Y_{ik'} \;=\; 1, Y_{ik} \;=\; 0$ forms an error if $k' < k$, while pattern $Y_{ik} \;=\; 1, Y_{ik'} \;=\; 0$ forms an error if $k' > k$. The H coefficient for item k is accordingly defined by

$$H_k \;=\; 1 \;-\; \frac{\sum\limits_{k'=1}^{k-1} P\{Y_{ik'} = 1,\, Y_{ik} = 0\} \;+\; \sum\limits_{k'=k+1}^{K} P\{Y_{ik} = 1,\, Y_{ik'} = 0\}}{\sum\limits_{k'=1}^{k-1} P\{Y_{ik'} = 1\}\, P\{Y_{ik} = 0\} \;+\; \sum\limits_{k'=k+1}^{K} P\{Y_{ik} = 1\}\, P\{Y_{ik'} = 0\}} .$$

Similarly, the H coefficient for the whole scale is based on the total number of errors:

$$H \;=\; 1 - \frac{\sum_{k<k'} P\{Y_{ik} = 1,\, Y_{ik'} = 0\}}{\sum_{k<k'} P\{Y_{ik} = 1\}\, P\{Y_{ik'} = 0\}} \;.$$

It is useful to note here that the scalability coefficients are completely determined by the tracelines $p_k(.)$ together with the probability distribution in the object population of the latent trait values θ_i. The h coefficients are higher, in general, when tracelines are steeper and when the object distribution has greater dispersion. The scalability coefficients should not be regarded as indicators of unidimensionality as such, but rather as indicators of how well the given set of items performs in assigning unidimensional scale values to objects from the population under study.

These notions can be adapted to our two-level design. We wish to know how well each object i can be measured by the set of responses Y_{ijk} ($k = 1, ..., K$; $j = 1, ..., n_i$). It is to be expected that responses given by the same subject are more consistent, i.e., contain less errors, than responses given by different subjects. Accordingly, we distinguish *within-subject scalability coefficients* from *between-subject scalability coefficients*, where the former are expected to be higher than the latter.

For item pairs $k < k'$, the within-subject scalability coefficient is defined as

$$H^W_{kk'} \;=\; 1 - \frac{P\{Y_{ijk}=1,\, Y_{ijk'}=0\}}{P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}}\;, \tag{4}$$

where attention must be given the fact that a single subject j is considered. The between-subject scalability coefficient is defined as

$$H^B_{kk'} \;=\; 1 - \frac{P\{Y_{ijk}=1,\, Y_{ij'k}=0\}}{P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}}\;\; (j \neq j')\;, \tag{5}$$

where two different subjects j and j' are considered in the numerator. (Since the denominator is the product of two separate probabilities not depending on j, replacing the last j in the denominator by a j' would not make any difference.)

Analogous scalability coefficients for the items are the within-subject scalability coefficient

$$H^W_k \;=\; 1 - \frac{\sum\limits_{k'=1}^{k-1} P\{Y_{ijk'}=1,\, Y_{ijk}=0\} + \sum\limits_{k'=k+1}^{K} P\{Y_{ijk'}=0,\, Y_{ijk}=1\}}{\sum\limits_{k'=1}^{k-1} P\{Y_{ijk'}=1\}\, P\{Y_{ijk}=0\} + \sum\limits_{k'=k+1}^{K} P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}} \tag{6}$$

and the between-subject scalability coefficient

$$H^B_k \;=\; 1 - \frac{\sum\limits_{k'=1}^{k-1} P\{Y_{ij'k'}=1,\, Y_{ijk}=0\} + \sum\limits_{k'=k+1}^{K} P\{Y_{ij'k'}=0,\, Y_{ijk}=1\}}{\sum\limits_{k'=1}^{k-1} P\{Y_{ijk'}=1\}\, P\{Y_{ijk}=0\} + \sum\limits_{k'=k+1}^{K} P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}}. \tag{7}$$

For the entire scale, the within-subject scalability coefficient is defined by

$$H^W \;=\; 1 - \frac{\sum_{k<k'} P\{Y_{ijk}=1,\, Y_{ijk'}=0\}}{\sum_{k<k'} P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}}, \tag{8}$$

and the between-subject scalability coefficient by

$$H^B \;=\; 1 - \frac{\sum_{k<k'} P\{Y_{ijk}=1,\, Y_{ij'k'}=0\}}{\sum_{k<k'} P\{Y_{ijk}=1\}\, P\{Y_{ijk'}=0\}}\;\; (j \neq j'). \tag{9}$$

The *within-subject* scalability coefficients refer to the situation that the two levels are combined (or 'disaggregated') by considering every object-subject combination as a single case. In other words, they are the usual H coefficients for scalability of the items for object-subject combinations treated as independent replications. The probabilities in the definitions of the within-subject scalability coefficients are determined by the tracelines $p_k(.)$ and the distribution of the latent parameter values $\theta_i + \delta_{ij}$, e.g.,

$$P\{Y_{ijk}=1,\, Y_{ijk'}=0\} \;=\; E_\theta E_\delta \left\{ p_k(\theta_i+\delta_{ij})(1-p_{k'}(\theta_i+\delta_{ij}))\right\}\;. \tag{10}$$

The *between-subject* scalability coefficients, on the other hand, are coefficients for scalability of the objects when the items are responded to by different, i.e., independent subjects. The probabilities in their definitions are determined by the tracelines $\pi_k(.)$ given in (3) and the distribution of the objects' latent trait values θ_i, as is expressed in

$$P\{Y_{ijk} = 1, Y_{ij'k'} = 0\} \; = \; E_\theta \left\{\pi_k(\theta_i)(1 - \pi_{k'}(\theta_i))\right\}. \tag{11}$$

Since the tracelines $\pi_k(.)$ are flatter than $p_k(.)$, and the distribution of $\theta_i + \delta_{ij}$ is more dispersed than that of θ_i, the within-subject scalability coefficients must be greater than between-subject scalability coefficients, unless the deviations δ_{ij} are constant, in which case they are equal. This is indeed proven in the appendix. Thus, it holds that $0 \leq H^B \leq H^W$. Suppose that $\text{var}(\theta_i + \delta_{ij}) > 0$ and the tracelines are strictly increasing; then $H^W > 0$. In the extreme situation that $\text{var}(\delta_{ij}) = 0$, it holds that $0 < H^B = H^W$. In the extreme situation that $\text{var}(\theta_i) = 0$ we have $0 = H^B < H^W$, and, scaling of objects makes no sense.

For the investigation of the quality of the scale as a unidimensional cumulative scale for object-subject combinations, the within-subject scalability coefficients are the most relevant. Their interpretation is similar to the interpretation of scalability coefficients in the usual Mokken scaling model. In the usual Mokken scaling procedure, a value of 0.3 is considered a low value for scalability, while 0.5 is good (e.g., Mokken and Lewis, 1982). In the present two-level situation, however, within-subject scalability coefficients may be allowed to be lower than 0.3, because of the presence of several subjects per object.

For the investigation of the degree to which the object value on the latent trait, θ_i, determines the responses to the items, the between-subject scalability coefficients, and their relation to within-subject coefficients, provide useful information. The discussion given above about the relation between H^B and H^W implies that the ratio H^B/H^W can be used as an indication of between-to-within-subject variability of the latent parameter. When H^B is almost as large as H^W, then the responses for the object are hardly affected by the particular subject (the influence of the random deviation δ_{ij} is small); when, on the contrary, H^B is much smaller than H^W, then the responses for the object are strongly affected by the particular subject (the influence of δ_{ij} is large). Similar interpretations can be attached to the ratios $H^B_{kk'}/H^W_{kk'}$ and H^B_k/H^W_k. Relatively small values of the between-subject coefficients imply that a large number of subjects (or items) is necessary for a reliable estimation of object scores. Very small values of the between-subject coefficients suggest that the object parameter θ_i is perhaps not a very relevant parameter because its influence on the responses to the items is small compared to the influence of the subjects, expressed by the deviations δ_{ij}.

Another way to assess the effect of the random subjects is to compute within-object and between-object covariance matrices of the items. These covariance matrices are, like the scalability coefficients, based on the probabilities of the item responses and of joint responses to pairs of items. The within-object correlation, which can be computed from these covariance matrices, shall be considered in Section 4.

4 Estimation of the scalability coefficients

The various scalability coefficients can be estimated by taking the defining formulae amd substituting relative frequencies for the probabilities. Recall that the number of subjects presented to object i is n_i; define the total number of subjects presented by n_+, and the total number of objects by N. We assume that the frequencies n_i are stochastically independent of the random variables Y_{ijk}, conditional on the latent trait values θ_i and δ_{ij}; we shall see that it is not a big problem if the n_i are correlated with the latent trait values θ_i.

If the numbers n_i are not the same for all i, then there are two ways to estimate the probabilities in formulas (4) to (9). We discuss their difference for the probabilities $P\{Y_{ijk}=1\}$. The first way is to average the *relative frequencies* for all objects:

$$\hat{P}\{Y_{ijk}=1\} \;=\; \frac{1}{N}\sum_i \frac{1}{n_i}\sum_j Y_{ijk} \;. \tag{12}$$

The other way is to average the *frequencies* for the objects:

$$\hat{P}\{Y_{ijk}=1\} \;=\; \frac{1}{n_+}\sum_i \sum_j Y_{ijk} \;. \tag{13}$$

To understand the difference between these two estimators, note that the probability $P\{Y_{ijk}=1\}$ is defined by

$$P\{Y_{ijk}=1\} \;=\; E_\theta E_\delta \, P\{Y_{ijk}=1 \,|\, \theta_i,\, \delta_{ij}\} \;=\; E_\theta E_\delta \, p_j(\theta_i+\delta_{ij}) \;,$$

where the subscripts θ and δ indicate that the expectation is taken with respect to the corresponding random variables. The probability for object i,

$$P\{Y_{ijk}=1 \,|\, \theta_i\} \;=\; E_{\delta_{ij}} P\{Y_{ijk}=1 \,|\, \theta_i,\, \delta_{ij}\} \;,$$

is estimated unbiasedly by the relative frequency for object i,

$$\hat{P}\{Y_{ijk} = 1 \,|\, \theta_i\} \;=\; \frac{1}{n_i}\sum_j Y_{ijk} \;.$$

This demonstrates that (12) defines an unbiased estimator, even if there is a stochastic dependence between n_i and θ_i.

Since estimator (13) weighs the individual relative frequencies by n_i/n_+, the latter estimator is unbiased only if the "sample sizes" n_i and the latent trait values θ_i are stochastically independent. Since this assumption is not always warranted, we opt for estimator (12). This estimator shall be denoted $\hat{P}_k$.

The estimators for the probabilities of error patterns, analogous to (12), are

$$\hat{P}^W_{k,k'}(1,0) \;=\; \hat{P}\{Y_{ijk} = 1,\, Y_{ijk'} = 0\} \;=\; \frac{1}{N}\sum_i \frac{1}{n_i}\sum_j Y_{ijk}(1 - Y_{ijk'}) \qquad (14)$$

and

$$\hat{P}^B_{k,k'}(1,0) \;=\; \hat{P}\{Y_{ijk} = 1,\, Y_{ij'k'} = 0\} \;=\; \frac{1}{N}\sum_i \frac{1}{n_i(n_i-1)}\sum_{j\neq j'} Y_{ijk}(1-Y_{ij'k'}).(15)$$

Substitution of these estimators leads to the following estimators of the scalability coefficients:

$$H^W_{kk'} \;=\; 1 - \frac{\hat{P}^W_{k,k'}(1,0)}{\hat{P}_k(1-\hat{P}_{k'})} \;, \qquad (16)$$

$$H^B_{kk'} \;=\; 1 - \frac{\hat{P}^B_{k,k'}(1,0)}{\hat{P}_k(1-\hat{P}_{k'})} \;, \qquad (17)$$

$$H^W \;=\; 1 - \frac{\sum_{k<k'} \hat{P}^W_{k,k'}(1,0)}{\sum_{k<k'} \hat{P}_k(1-\hat{P}_{k'})} \;, \qquad (18)$$

$$H^B \;=\; 1 - \frac{\sum_{k<k'} \hat{P}^W_{k,k'}(1,0)}{\sum_{k<k'} \hat{P}_k(1-\hat{P}_{k'})} \;, \qquad (19)$$

and similarly for H^W_k and H^B_k.

As a final remark it can be added that if disaggregated data, i.e., data where object-subject combinations are treated as independent replications, are used as input in the MSP program for Mokken scaling, then the scalability coefficients computed are the *within*-subject coefficients, but with estimates analogous to (13), and therefore weighted by n_i/n_+. If all n_i are equal, then this is correct. If the n_i values are considerably different from each other, however, while they are correlated with the latent trait values θ_i, this approach can yield incorrect results.

5 Object scores

Object scores can be defined as

$$\overline{Y}_{i..} = \frac{1}{m\,n_i} \sum_{j,k} Y_{ijk}. \tag{20}$$

Each index over which has been averaged is replaced by a dot. Just like in the usual Mokken model, these object scores are not estimators for θ_i, because the non-parametric nature of the model makes it impossible to estimate these latent values: they are not identifiable. The expectation of $\overline{Y}_{i..}$ is

$$\mu(\theta_i) = \frac{1}{m} \sum_{k=1}^{m} \pi_k(\theta_i), \tag{21}$$

a monotone function of θ_i. So, apart from chance fluctuations, the relation between $\overline{Y}_{i..}$ and θ_i is monotone. Of course the standard error of estimation of $\overline{Y}_{i..}$ depends on n_i as well as on θ_i.

6 Reliability

For the usual Mokken scaling method, there are several methods to estimate reliability; these methods of estimation are based on the matrix of joint positive responses to pairs of items, the so-called P matrix, see Mokken (1971) and Sijtsma and Molenaar (1987). A nice feature of our two-level design is the availability of independent within-object replications. These replications imply that a within-object between-subject test-retest correlation coefficient can be estimated as the intra-class correlation coefficient, where classes are defined by the objects. Define σ_0^2 as the variance of average scores for randomly drawn object-subject combinations,

$$\sigma_0^2 = \mathrm{var}(\overline{Y}_{ij.}) \ .$$

Further define the within-object between-subject correlation

$$\rho = \mathrm{corr}(\overline{Y}_{ij.}, \overline{Y}_{ij'.}) = \frac{1}{\sigma_0^2}\,\mathrm{cov}(\overline{Y}_{ij.}, \overline{Y}_{ij'.}) \quad (j \neq j')\,;$$

then ρ can be regarded as the intra-class correlation coefficient, where classes are defined by the objects. This parameter will be called the *intra-object correlation coefficient.* The variance of the object score for a random object with n_i subjects is

$$\mathrm{var}\ (\overline{Y}_{i..}) \ = \ \frac{\sigma_0^2}{n_i}\,\{1 + (n_i - 1)\rho\} \ . \tag{22}$$

To define estimators for σ_0^2 and ρ, denote the average object score by

$$\overline{Y}_{...} = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_{i..} \ ;$$

note that $\overline{Y}_{...}$ is not the mean of all Y_{ijk} values, but the mean of the object scores $\overline{Y}_{i..}$; these two means are the same if all the n_i are equal. Further, denote between-object and within-object variances by

$$S_B^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

and

$$S_W^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (\overline{Y}_{ij.} - \overline{Y}_{i..})^2 \ .$$

(The weights $1/(n_i - 1)$ are used in order to let S_W^2 be the average of the observed within-object variances; since the object variances are not necessarily the same, this is the only way to get an unbiased estimator for the expectation of the theoretical covariance for a randomly drawn object.) Define the harmonic mean ν of the n_i by

$$\frac{1}{\nu} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \ .$$

It can be proved straightforwardly that

$$E(S_W^2) = \sigma_0^2 (1 - \rho)$$

and

$$E(S_B^2) = \frac{\sigma_0^2}{\nu} \{1 + (\nu - 1)\rho\} \ .$$

This yields the estimators

$$\hat{\sigma}_0^2 \ = \ \frac{\nu - 1}{\nu} S_W^2 \ + \ S_B^2 \qquad (23)$$

$$\hat{\rho} \ = \ \frac{1}{\hat{\sigma}_0^2} \{S_B^2 \ - \ \frac{1}{\nu} S_W^2\} \ . \qquad (24)$$

The estimated intra-object correlation $\hat{\rho}$ is the test-retest reliability for the parallel tests corresponding to the different subjects. The test-retest correlation for two independent scores $\overline{Y}_{i..}$ for the same object i (obtained from two independent sets each consisting of n_i independent subjects) is the well-known reliability coefficient known as coefficient alpha (see, e.g., Nunnally, 1967, p. 193)

$$\rho(\overline{Y}_{i..}, \overline{Y}'_{i..}) = \frac{n_i \rho}{(n_i - 1)\rho + 1} = \frac{n_i \rho}{n_i \rho + 1 - \rho} . \tag{25}$$

This reliability coefficient will be very important in the final assessment of the quality of the scale for objects, as an addition to the Loevinger scalability coefficients treated in Section 3.

7 Examples for simulated data

The presentation of some results for simulated data may give a better understanding of the various parameters introduced and the numerical values these may assume. Three data sets were generated for $N = 500$ objects, each confronted with n_i= 10 subjects for which $K = 6$ items were scored. The tracelines of the items satisfied the Rasch model (see, e.g., Fischer and Molenaar, 1995):

$$p_k(\theta) = \frac{\exp(\theta - \xi_k)}{1 + \exp(\theta - \xi_k)} ,$$

where ξ_k is the difficulty parameter of item k. The difficulty parameters were chosen as -2.0, -1.2, -0.4, 0.4, 1.2, 2.0. The distributions of θ_i and δ_{ij} were normal with mean 0.

The three simulated data sets differed according to the ratio of within-subject to between-subject variances: the variances were $\sigma^2_\theta = 0.5$, $\sigma^2_\delta = 0.5$ (simulation 1), $\sigma^2_\theta = 0.8$, $\sigma^2_\delta = 0.2$ (simulation 2), and $\sigma^2_\theta = 0.2$, $\sigma^2_\delta = 0.8$ (simulation 3). Thus, the variance of the combined latent trait values $\theta_i + \delta_{ij}$ is in all three cases equal to 1. This implies that the item popularities (marginal probabilities of a positive response) and the within-subject H coefficients are the same in the three simulations.

For simulation 1, where the variance of object values for the latent trait is equal to the variance of the subject-related deviations, the popularities and scalability coefficients were estimated as presented in Tables 1, 2 and 3.

We see, what is well known for Mokken scaling, that H coefficients are greater when the item popularities are further apart. Further, the between-subject H coefficients are indeed considerably smaller than the within-subject coefficients.

Table 1: Estimated item probabilities for simulated data set 1.

k	1	2	3	4	5	6
$\hat{P}_k$	0.158	0.266	0.409	0.577	0.734	0.842

Table 2: Estimated within-subject scalability coefficients
for simulated data set 1.

$H^W_{kk'}$	k'	1	2	3	4	5	6
k							
1		—	0.138	0.210	0.349	0.411	0.432
2			—	0.155	0.283	0.309	0.452
3				—	0.212	0.290	0.401
4					—	0.221	0.278
5						—	0.182
6							—
Items							
k		1	2	3	4	5	6
H^W_k		0.253	0.227	0.231	0.253	0.256	0.301
Whole scale							
H^W		0.250					

Table 3: Estimated between-subject scalability coefficients for simulated data set 1.

$H^B_{kk'}$	k'	1	2	3	4	5	6
k							
1		—	0.067	0.098	0.195	0.164	0.182
2			—	0.078	0.137	0.147	0.188
3				—	0.113	0.104	0.135
4					—	0.086	0.111
5						—	0.075
6							—
Items							
k		1	2	3	4	5	6
H^B_k		0.121	0.108	0.102	0.120	0.103	0.118
Whole scale							
H^B		0.111					

Table 4: Estimated values for coefficient alpha for simulated data set 1.

n_i	4	8	10	20	30
$\rho(\overline{Y}_{i..}, \overline{Y}_{i..}{}')$	0.549	0.709	0.753	0.859	0.901

Some further parameter estimates are $\hat{\sigma}_0^2 = 0.052$, $\hat{\rho} = 0.233$. Some values for reliability coefficient alpha are given in Table 4.

For this data set, the within-subject coefficients are rather low compared to the value of 0.3 that is considered a reasonable minimum for one-level Mokken scaling. The ratio H^B/H^W is 0.443. We see that we need at least 10 to 20 subjects per object in order to get a sufficiently reliable scale.

For the second data set, where the influence of the subjects is much smaller ($\sigma_\theta^2 = 0.8$, $\sigma_\delta^2 = 0.2$), we obtained $H^W = 0.260$, $H^B = 0.211$, and therefore a high value $H^B/H^W = 0.809$; further, $\hat{\sigma}_0^2 = 0.052$ and $\hat{\rho} = 0.377$. Note that the population values for H^W and for σ_0^2 are here the same as for the first simulation. Table 5 gives values for the reliability coefficient.

Table 5: Estimated values for coefficient alpha for simulated data set 2.

n_i	4	8	10	20	30
$\rho(\overline{Y}_{i..}, \overline{Y}'_{i..})$	0.708	0.829	0.858	0.924	0.945

It is seen that much smaller numbers of subjects suffice for obtaining a reliable scale than for simulation 1. For the third simulation, where the influence of subjects is much greater, we obtain the reverse results. Some parameter estimates are $H^W = 0.258$, $H^B = 0.058$; and further $\hat{\sigma}_0^2 = 0.053$, $\hat{\rho} = 0.106$. Table 6 gives values for the reliability coefficient.

Table 6: Estimated values for coefficient alpha for simulated data set 3.

n_i	4	8	10	20	30
$\rho(\overline{Y}_{i..}, \overline{Y}'_{i..})$	0.321	0.487	0.542	0.703	0.781

For the parameters $\sigma_\theta^2 = 0.2$, $\sigma_\delta^2 = 0.8$ of this simulation, the relative effect of the subjects is so large that it is doubtful whether it makes much sense to scale objects on the basis of this set of items and this population of subjects. This is reflected by the low value 0.226 of H^B/H^W, and by the low reliabilities.

Some tentative conclusions may be drawn regarding desirable values of the various parameters. The fact that multiple parallel measurements are available allows the use of scales with lower H values than in single-level nonparametric scaling still having a satisfactory reliability. Within-subject scalability coefficients for items and for the whole scale should be greater than 0.2 for a good scale; it is not serious if for some directly consecutive item pairs, the $H^W_{kk'}$ coefficient is between 0.1 and 0.2. The consistency between subjects is satisfactory when between-subject homogeneity coefficients are at least 0.1 (with possibly some exceptions for between-item $H^B_{kk'}$ coefficients). For the ratio H^B/H^W, values over 0.3 are reasonable and values over 0.6 are excellent. As indicators for the quality of the scale one should use not only the homogeneity coefficients, but also coefficient alpha for practical n_i values.

8 Example: assessment of teachers by pupils

In a study by Bosker and others (1999), pupils in primary schools were asked to respond to a questionnaire about their teacher and classroom. As an example, some questions about the classroom climate and order are used. Six items from the questionnaire were selected on the basis of face value considerations (mainly their unambiguous relation to order in the class). The items are statements with three answer categories ('true', 'somewhat true', 'not true'). They were recoded to values 1, 2, 3, where 1 denotes the most orderly and 3 the most chaotic situation in the classroom. A cross-validatory approach was chosen in which the first phase used data for group-6 pupils (age 9–10 years) for the selection of a good subset of items and dichotomization thresholds, and the second phase used the data for group-7 pupils (age 10–11 years) for the evaluation of the resulting scale along the lines of the preceding sections.

In the first phase, items and thresholds were chosen – in a trial and error procedure – so as to yield relatively high between-subject H coefficients for each item. This resulted in a scale of four items, all dichotomized by contrasting the values 1 and 2 (most and intermediate orderly, coded as $Y = 0$) with the value 3 (least orderly, coded as $Y = 1$). The four items are the following. They are ordered in increasing frequency of the least orderly outcome.

1. When the teacher tells us something, we listen well (inversely coded).

2. It is usually quiet in the classroom (inversely coded).

3. There often is much noise in the classroom.

4. The teacher often tells us to be quiet.

These items were subjected to a scale analysis for the group-7 pupils (whose data were not used in the selection of items and thresholds) as the second phase of the data analysis. There were data for 1530 pupils (subjects) and 77 classrooms/teachers (objects). The results are presented in Tables 7 to 10.

Table 7: Estimated item probabilities for classroom climate scale.

k	1	2	3	4
$\hat{P}_k$	0.062	0.277	0.351	0.524

Table 8: Estimated within-subject scalability coefficients for classroom climate scale.

$H^W_{kk'}$	k'	1	2	3	4
k					
1		—	0.697	0.740	0.676
2			—	0.562	0.555
3				—	0.556
4					—
Items					
k		1	2	3	4
H^W_k		0.707	0.576	0.578	0.566
Whole scale					
H^W		0.587			

The between-to-within subject ratio of the scalability coefficient for the entire scale is $H^B/H^W = 0.367$. The intra-object correlation is $\hat{\rho} = 0.258$.

It can be concluded that this four-item scale is very satisfactory. Within-subject pairwise H coefficients are 0.55 and higher, between-subject pairwise H coefficients are 0.18 and higher. The consistency of response within and also between subjects is large enough to use these four questions for assessing classroom

Table 9: Estimated between-subject scalability coefficients for classroom climate scale.

$H^B_{kk'}$	k'	1	2	3	4
k					
1		—	0.324	0.324	0.396
2			—	0.179	0.204
3				—	0.201
4					—
Items					
k		1	2	3	4
H^B_k		0.343	0.207	0.204	0.220
Whole scale					
H^B		0.215			

Table 10: Estimated values for coefficient alpha for for classroom climate scale.

n_i	4	8	10	20	30
$\rho(\overline{Y}_{i..}, \overline{Y}^{'}_{i..})$	0.582	0.736	0.777	0.874	0.913

climate. About 10 pupils are sufficient to give a reasonably reliable measurement instrument.

9 Discussion

A non-parametric method has been presented to scale objects who are known to the researcher only through responses which refer to subjects connected in some way to the objects. The subjects, nested within the objects, are supposed to be a random sample from a population, and a test composed of dichotomous items is administered to each object-subject combination. This makes for a two-level data structure.

The consistency of answer patterns within subjects can be measured by within-subject Loevinger H coefficients applied to the disaggregated object-subject combinations. To scale the objects on the basis of such data requires, however, that there is enough consistency also *between* the subjects in their answer patterns. This is measured by between-subject Loevinger H coefficients.

To investigate how well the objects are scaled by this set of items, two measures have been presented: the ratio H^B/H^W of *between-subject* (within-object) to within-subject scalability coefficients, which gives an indication of the extent to which scale values are determined by objects rather than by subjects (and object-subject interactions); and coefficient alpha for the reliability of the resulting scale. Coefficient alpha depends, of course, on the number of subjects per object.

Scales for objects defined in this way can be used, e.g., to investigate relations between the corresponding latent trait and other variables referring to the objects, such as the relation between classroom climate and teacher behavior, or the relation between neighborhood climate and policy instruments applied to the neighborhoods.

A possible alternative to the non-parametric approach presented here is a parametric three-level model for dichotomous data, where the levels are items, subjects, and objects. A standard approach, elaborated for the three-level case by Gibbons and Hedeker (1997), is to assume normal distributions for the latent trait components θ_i and δ_{ij}, and a logistic link function for the dichotomous responses. Although this may be called a standard approach from the point of view of model building, the numerical calculations necessary to estimate the parameters are very complicated. Reviews of parametric models for multilevel dichotomous data (but with a focus on two-level models) are given by Goldstein

(1995, Chapter 7) and Snijders and Bosker (1999, Chapter 14). Such an approach yields the rewards connected to the richer parametric structure: the possibility of using covariates for subjects and/or objects, the availability of standard statistical methods such as maximum likelihood to obtain estimates and standard errors. However, for these models the estimation methods are numerically quite complex and computationally demanding. Moreover, the parametric assumptions will often be questionable; the logistic link function means that one assumes the Rasch model for the object-subject combinations, and the Rasch model is known to be quite a strong assumption for a cumulative unidimensional scale.

Advantages of the non-parametric method presented here are the light assumptions and the easy computations. A PC program to carry out the calculations is available from the web site http://stat.gamma.rug.nl/snijders/multilevel.htm .

A Appendix. Proof that between-subject scalability coefficients are not larger than within-subject coefficients.

It shall be proved that

$$P\{Y_{ijk} = 1, Y_{ijk'} = 0\} \le P\{Y_{ijk} = 1, Y_{ij'k'} = 0\} \quad (j \neq j') \,. \tag{26}$$

This is equivalent with $H^B_{kk'} \le H^W_{kk'}$, and it implies $H^B_k \le H^W_k$ and $H^B \le H^W$. To prove (26), note that

$$\begin{aligned} & P\{Y_{ijk} = 1,\, Y_{ij'k'} = 0\} \,-\, P\{Y_{ijk} = 1,\, Y_{ijk'} = 0\} \\ & \quad = \,(P\{Y_{ijk} = 1\} \,-\, P\{Y_{ijk} = 1,\, Y_{ij'k'} = 1\}) \\ & \qquad - \,(P\{Y_{ijk} = 1\} \,-\, P\{Y_{ijk} = 1,\, Y_{ijk'} = 1\}) \\ & \quad = \,P\{Y_{ijk} = 1,\, Y_{ijk'} = 1\} \,-\, P\{Y_{ijk} = 1,\, Y_{ij'k'} = 1\} \;. \end{aligned}$$

From the expressions analogous to (10) and (11) we can conclude that this equals

$$\begin{aligned} & E_\theta \, E_\delta \, \{p_k(\theta_i + \delta_{ij}) p_{k'}(\theta_i + \delta_{ij})\} \,-\, E_\theta \, \{\pi_k(\theta_i)\pi_{k'}(\theta_i)\} \\ & \quad = \,E_\theta \, [E_\delta \, \{p_k(\theta_i + \delta_{ij}) p_{k'}(\theta_i + \delta_{ij}) \,|\, \theta_i\} \,-\, \pi_k(\theta_i)\pi_{k'}(\theta_i)] \;. \end{aligned}$$

Because of the definition of π_k, given in (3), this is equal to

$$E_\theta \, [\mathrm{cov}_\delta\{p_k(\theta_i + \delta_{ij}),\, p_{k'}(\theta_i + \delta_{ij}) \,|\, \theta_i\}] \,.$$

The conditional covariance,

$$\mathrm{cov}_\delta\{p_k(\theta_i+\delta_{ij}),\, p_{k'}(\theta_i+\delta_{ij})\,|\,\theta_i\}\ ,$$

is a covariance of two non-decreasing functions of the random variable δ_{ij}. Such a covariance is always non-negative, which proves (26). If the tracelines $p_k(.)$ are strictly increasing and $\mathrm{var}(\delta_{ij}) > 0$, then this covariance is strictly positive, so that the strict inequality $H^W_{kk'} > H^B_{kk'}$ holds.

In a similar way it can be proved that

$$\begin{aligned}&P\{Y_{ijk}=1, Y_{ij'k'}=0\} - P\{Y_{ijk}=1\}P\{Y_{ijk'}=0\}\\ &\quad = -\,\mathrm{cov}_\theta\,\{\pi_k(\theta_i),\, \pi_{k'}(\theta_i)\} \le 0\ ,\end{aligned}$$

which implies $H^B_{kk'} \ge 0$.

References

Bosker, R.J., and others (1999), *Zelfevaluatie in het basisonderwijs (ZEBRA).* University of Twente.

Fischer, G.H., and Molenaar, I.W. (1995), *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer.

Gibbons, R.D., and Hedeker, D. (1997), Random effects probit and logistic regression models for three-level data. *Biometrics*, 53, 1527–1537.

Goldstein, H. (1995), *Multilevel Statistical Models*, 2nd ed. London: Edward Arnold.

Jansson, I., and Spreen, M. (1998), The use of local networks in a study of heroin users: assessing average local networks. *Bulletin de Méthodologie Sociologique*, 59, 49–61.

Mokken, R.J. (1971), *A theory and procedure of scale analysis.* The Hague: Mouton.

Mokken, R.J. (1971), Nonparametric models for dichotomous responses, pp. 351–367 in W.J. van der Linden and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*, New York: Springer.

Mokken, R.J., and Lewis, C. (1982), A nonparametric approach to the analysis of dichotomous item responses, *Applied Psychological Measurement*, 6, 417–430.

Molenaar, W., and Sijtsma, K. (200?). Textbook on non-parametric scaling.

Nunnally, J.C. (1967), *Psychometric Theory.* New York: MacGraw-Hill.

Raudenbush, S.W., and Bryk, A.S. (1992), *Hierarchical Linear Models*, Newbury Park: Sage.

Raudenbush, S.W., and Sampson, R.J. (1998), "Ecometrics": Toward a science of assessing ecological settings, with applications to the systematic social observatiopn of neigborhoods. Paper under review.

Sijtsma, K., and Molenaar, I.W. (1987), Reliability of test scores in nonparametric item response theory, *Psychometrika*, 52, 79–97.

Snijders, T.A.B., and Bosker, R.J. (1999), *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.